IDS 572 Assignment 2 Case– Loan default prediction and investment strategies in online lending

Due date: Oct 26, 2022

In this assignment, we will analyze data from an online lending platform, Lending Club. The goal is to develop models to predict which loans are at risk of default. Such models can then be used to devise investment strategies.

Background

P2P lending platforms - like Lending Club (LC), Prosper, Peerform, Upstart, etc - provide an online environment for matching borrowers seeking loans and lenders looking to make an investment. With lower operational costs than traditional lenders (banks), such online lending platforms leverage technology, data and analytics to bring quicker and more convenient financing for individual and small business borrowers from investors looking for attractive investment yields. With increasing volumes, what started as peer-to-peer platforms for connecting individual borrowers and individual investors has today evolved to include institutional investors, hedge funds, etc. Also called marketplace lending or alternate lending, such fintech platforms have seen significant growth in recent years. It is estimated that in 2018, 38% of all personal loans in the US were issued through fintech firms, growing from 5% in 2013 [11]. Some estimate the global online lending market to grow from ~$42B in 2018 to ~$460B in 2022 2[2]. Lending Club, a pioneer in fintech, is one of the largest online lending platforms, with over $50B in total loans issued till date 3[3]. (See a comparison at https://www.investopedia.com/articles/investing/092315/7-best-peertopeer-lending-websites.asp)

"LendingClub uses technology to operate its online credit marketplace at a lower cost than traditional lending programs, passing the savings on to borrowers in the form of lower rates, and offering investors the potential for competitive returns"[4] Further information is detailed in their website, which you should examine to understand how borrowers apply for loans and the information available for investors to decide on loans to finance.

Lending Club has very recently (Jan 2021) received approval for acquiring the digital bank, Raduis, and will stop operations in its current P2P platform. As reported in Jan 19, 2021:
"This is a transformative acquisition for the company and a watershed moment for the industry as we become the only full-spectrum fintech marketplace bank in the U.S.," said Scott Sanborn, CEO of LendingClub. "The customer benefits of this acquisition are even clearer now that COVID has accelerated Americans' move to digital banking. As the only full-spectrum fintech marketplace bank, LendingClub will be able to use our technology and data-driven platform to provide new products and services to our millions of members that will help them both pay less when borrowing and earn more when saving. By combining with Radius, we will create a category-defining experience that will also dramatically enhance the resilience and earnings trajectory of our business."
(https://www.prnewswire.com/news-releases/lendingclub-receives-regulatory-approvals-to-acquire-radius-bancorp-301210498.html)

[1] https://www.cnbc.com/2019/02/21/personal-loans-surge-to-a-record-138-billion-in-us-as-fintechs-lead-new-lending-charge.html
[2] https://www.alliedmarketresearch.com/peer-to-peer-lending-market
[3] https://www.lendingclub.com/info/statistics.action
[4] https://www.lendingclub.com/public/how-peer-lending-works.action

This assignment is based on data from Lending Club (LC). Similar data is available from other P2P lending platforms, to help investors assess riskiness of loans and make investment decisions. LC issues personal loans between $1000 and $40,000 for 36 to 60 month durations. Interest rates on these loans are determined based on a variety of information, including credit rating, credit history, income, etc. Based on this, LC assigns a grade for each loan, ranging from A for safest loans to G for highest risk; subgrades are also assigned within each grade. Loans are split into $25 notes, which investors can purchase. Interested investors can browse different loans the LC website, which shows the assigned loan grade and other information.

The online lending business model and how Lending Club operates is described in various web resources. Having an understanding of this is important, to appreciate the role of data and analytics, and the future potential of this rapidly developing area of fintech.

An introduction to alternative lending. Morgan Stanley Investment Insights, May, 2019
https://www.morganstanley.com/im/en-us/financial-advisor/insights/investment-insights/an-introduction-to-alternative-lending.html
https://en.wikipedia.org/wiki/LendingClub
A Trillion Dollar Market By the People, For the People – How Marketplace Lending Will Remake Banking As We Know https://foundationcapital.com/wp-content/uploads/2020/04/FC_CharlesMoldow_TrillionDollarMarket.pdf
You may find this interesting - "Theorem uses data science and machine learning to invest in marketplace lending loans" https://www.theoremlp.com/
"LendingClub (A): Data Analytic Thinking" Harvard Business School Case, 2018. Our work in this assignment, though sharing some aspects of this case, takes a different approach in analyses.

To facilitate investment, P2P lenders provides access to their data. Large sets of data are provided in different files. For the purpose of this assignment, we will use a sample of loans issued during 2013-2015. The data carries information on 36 month loans, which will all have completed their term by now. Some loans were fully paid back, while others were "charged off" (defaulted).

Assignment

The data on loans is in the file lcDataSampleF22.csv and the LCDataDictionalry.xls file describes the variables.

In the first part of this assignment, we will explore the data on loans, to develop an understanding of loan grades and subgrades and how they may relate to default and returns performance, loan purpose and any relation to performance, analyses of returns from loans, etc. We also need to look into missing data, and how to address this. While the data carries information on over 100 variables, we need to determine *which data will be available when looking to invest in a loan* — since our goal is to develop a model to predict loan default and then decide which loans to invest in; such a model will thus be only able to consider variables available *before* a loan is issued.

The subsequent task is to develop models to identify good/bad loans ('fully paid' or 'charged off', and evaluate these. We will also consider investment performance corresponding to these models and identify the best model.

Questions:

Part A

1. (a) Your team's ultimate goal is to help clients determine whether they should invest in p2p loans. What is the final decision that you will help the client make? What is the objective, and how will you evaluate 'better' vs 'worse' decisions? What is the goal of predictive models for this ? What will be the potential target variables?

   (b) Take a look at the data attributes. How would you categorize these attributes, in broad terms, considering what they pertain to?
   Before doing any analyses, what do you think may be the important attributes to consider for your decision task?

2. Data exploration
   (a) some questions to consider:
   (i) What is the proportion of defaults ('charged off' vs 'fully paid' loans) in the data?
   How does default rate vary with loan grade? Does it vary with sub-grade? And is this what you would expect, and why?

   (ii) How many loans are there in each grade? And do loan amounts vary by grade?
   Does interest rate for loans vary with grade, subgrade? Look at the average, standard-deviation, min and max of interest rate by grade and subgrade. Is this what you expect, and why?

   (iii) For loans which are fully paid back, how does the time-to-full-payoff vary? For this, calculate the 'actual term' (issue-date to last-payment-date) for all loans. How does this actual-term vary by loan grade (a box-plot can help visualize this).

   (iv) Calculate the annual return. Show how you calculate the percentage annual return.
   Is there any return from loans which are 'charged off'? Explain. How does return from charged - off loans vary by loan grade?
   Compare the average return values with the average interest-rate on loans – do you notice any differences, and how do you explain this?
   How do returns vary by grade, and by sub-grade.
   If you wanted to invest in loans based on this data exploration, which loans would you invest in?

   (v)What are people borrowing money for (purpose)? Examine how many loans, average amounts, etc. by purpose? Do loan amounts vary by purpose? Do defaults vary by purpose? Does loan-grade assigned by Lending Club vary by purpose?

   (vi) Consider some borrower characteristics like employment-length, annual-income, fico-scores (low, high). How do these relate to loan attributes like, for example, loan_amout, loan_status, grade, purpose, actual return, etc.

   (vii) Generate some (at least 3) new derived attributes which you think may be useful for predicting default., and explain what these are. For these, do an analyses as in the questions above (as reasonable based on the derived variables).

(b) Are there missing values? What is the proportion of missing values in different variables? Explain how you will handle missing values for different variables. You should consider what he variable is about, and what missing values may arise from – for example, a variable monthsSinceLastDeliquency may have no value for someone who has not yet had a delinquency; what is a sensible value to replace the missing values in this case?
Are there some variables you will exclude from your model due to missing values?

(c) Consider the potential for data leakage. You do not want to include variables in your model which may not be available when applying the model; that is, some data may not be available for new loans before they are funded. Leakage may also arise from variables in the data which may have been updated during the loan period (ie., after the loan is funded). Identify and explain which variables will you exclude from the model.

3. Do a univariate analyses to determine which variables (from amongst those you decide to consider for the next stage prediction task) will be individually useful for predicting the dependent variable (loan_status). For this, you need a measure of relationship between the dependent variable and each of the potential predictor variables. Given loan-status as a binary dependent variable, which measure will you use? From your analyses using this measure, which variables do you think will be useful for predicting loan_status?
(Note – if certain variables on their own are highly predictive of the outcome, it is good to ask if this variable has a leakage issue).

We will next develop predictive models for loan_status.

4. (a) Split the data into training and validation sets. What proportions do you consider, why?
(b) How will you evaluate performance – which measure do you consider, and why?
For evaluation of models, you should include confusion matrix related measures, as well as ROC analyses and lifts. <u>Explain which performance measures you focus on, and why</u>.

5. Develop a decision tree model to predict default.
Train decision tree models (use either rpart or c50)
What parameters do you experiment with, and what performance do you obtain (on training and validation sets)? Clearly tabulate your results and briefly describe your findings.
[If something looks too good, it may be due to leakage – make sure you address this]

Identify the best tree model. Why do you consider it best?
Describe this model – in terms of complexity (size).
Examine variable importance. How does this relate to your uni-variate analyses in Question 3 above?
Briefly *describe* how variable importance is obtained (the process used in the decision tree learning algorithm you use(rpart or c50).

6. (a) Develop random forest and boosted tree model (using gbm or xgb)
Note the 'ranger' library and xgb can give faster computations.

What parameters do you experiment with, and how does this affect performance?
Describe the best random forest and boosted tree model in terms of number of trees, performance, variable importance.

(b)Compare the performance of random forest, boosted tree and decision tree model from Q 5 above. Do you find the importance of variables to be different ?
Which model would you prefer, and why ?

7. The purpose of the model is to help make investment decisions on loans. How will you evaluate the models on this business objective? Consider a simplified scenario - for example, that you have $100 to invest in each loan, based on the model's prediction. So, you will invest in all loans that are predicted to be 'Fully Paid'. Key questions here are: *how much, on average, can you expect to earn after 3 years from a loan that is paid off*, and *what is your potential loss from a loan that has to be charged off* ?

   One can consider the average interest rate on loans for expected profit – is this a good estimate of your profit from a loan? For example, suppose the average int_rate in the data is 11.2%; so after 3 years, the $100 will be worth (100 + 3*11.2) = 133.6, i.e a profit of $33.6. Now, is 11.2% a reasonable value to expect – what is the return you calculate from the data? Explain what *value of profit* you use.

   For a loan that is charged off, will the loss be the entire invested amount of $100? The data shows that such loans have do show some partial returned amount. Looking at the returned amount for charged off loans, what proportion of invested amount can you expect to recover? Is this overly optimistic? Explain which *value of loss* you use.

   You should also consider the alternate option of investing in, say in bank CDs (certificate of deposit); let's assume that this provides an interest rate of 2%. Then, if you invest $100, you will receive $106 after 3 years (not considering reinvestments, etc), for a profit of $6.
   Considering a confusion matrix, we can then have profit/loss amounts with each cell, as follows:

		Predicted	
		FullyPaid	ChargedOff
Actual	FullyPaid	*profitValue*	$6
	ChargedOff	*lossValue*	$6

   (a) Compare the performance of your models from Questions 5, 6 above based on this. Note that the confusion matrix depends on the classification threshold/cutoff you use. Which model do you think will be best, and why.

   (b) Another approach is to directly consider how the model will be used – you can order the loans in descending order of prob(fully-paid). Then, you can consider starting with the loans which are most likely to be fully-paid and go down this list till the point where overall profits

begin to decline (as discussed in class). Conduct an analyses to determine what threshold/cutoff value of prob(fully-paid) you will use and what is the total profit from different models – decision tree, random forest, boosted trees. Also compare the total profits from using a model to that from investing in the safe CDs. Explain your analyses and calculations.
Which model do you find to be best and why. And how does this compare with what you found to be best in part (a) above.

Part B: predictive models for loans with high returns

8. Develop models to identify loans *which provide the best returns*. Explain how you define returns? Does it include Lending Club's service costs?

   Develop glm, rf, gbm (xgb) models for this. Show how you systematically experiment with different parameters to find the best models. Compare model performance.

9. Considering results from the best model for predicting loan-status and that for predicting loan returns, how would you select loans for investment? There can be multiple approaches for combining information from the two models to make investment decisions (as d iscussed in class)– describe your approach, and show performance. How does performance here compare with use of single (i.e for predicting loan-status, or loan returns) models?

10. As seen in data summaries and your work in the first assignment, higher grade loans are less likely to default, but also carry lower interest rates; many lower grad loans are fully paid, and these can yield higher returns. Considering this, one approach to making investment decisions may be to focus on lower grade loans (C and below), and try to identify those which are likely to be paid off. Develop models from the data on lower grade loans, and check if this can provide an effective investment approach. Compare performance of models from different methods (glm, gbm, rf).
    Can this provide a useful approach for investment? Compare performance with that in Q9 above?

Please submit a pdf file with answers to the assignment questions, and supporting analyses. Also include a single Rmd file with your R code. Note – the code needs to be adequately commented and divided into sections in the Rmd file to help readability and ease understanding by others; arrange the Rmd file sections based on the assignment questions.